UCTNet: Uncertainty-aware Cross-modal Transformer Network for Indoor RGB-D Semantic Segmentation

Supplementary Material

Xiaowen Ying and Mooi Choo Chuah

Lehigh University
xiy517@lehigh.edu, chuah@cse.lehigh.edu

A.1 Additional Architecture Details

As we introduced in the main paper, both the RGB and depth encoders in our framework consist of a patch embedding layer and four sequential transformer blocks. The patch embedding layer first partitions the inputs into small patches of size 4×4 and uses a linear embedding layer to map each patch to a high-dimensional embedding. This layer quickly reduces the feature resolution to $\frac{H}{4} \times \frac{W}{4}$ similar to the Stem layers in the ResNets.

Each of the following Transformer Blocks is comprised of a Patch Merging layer (except block 1) and multiple Transformer Encoder Layers. The Patch Merging Layer is used to reduce the feature resolution by a factor of 4. The detail of the Transformer Encoder Layer is described in Section A.1.1.

A.1.1 Detail of the Transformer Encoders Layers.

The main difference between the architecture of the RGB encoder and the depth encoder is that the RGB encoder uses the Normal Transformer Block while the depth encoder employs the Uncertainty-Aware Transformer Blocks. The uncertainty-aware transformer block is constructed by replacing the Self-Attention layer in its encoder layers with our proposed Uncertainty-Aware Self-Attention layer. Figure A.1 shows the detailed architectures of the original Swin Encoder Layer and our Uncertainty-Aware Encoder layer. The window size M is set to 7 following the default configuration in [1].

A.2 Additional Ablation Study.

A.2.1 Varying the Temperature in Different Layers.

Our final framework uses a fixed temperature T (Equation 5) in every UASA layer for we found this setting is simple and effective. However, it is possible to let each UASA layer has its own temperature. In this section, we present two experiments that we have explored.

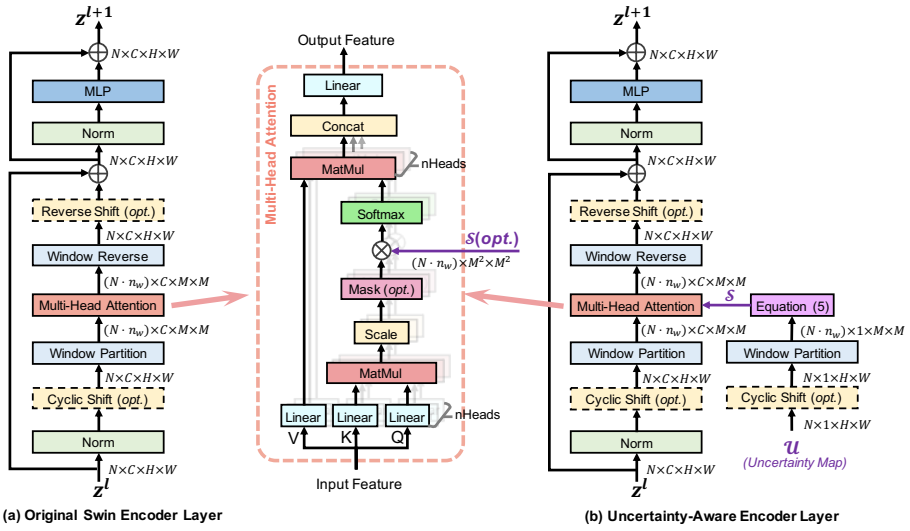

Fig. A.1. The detail architecture of the Uncertainty-Aware Encoder Layer compared to the original Swin Encoder Layer. [1]. Note that the shifting operations are only performed in the even layers. The relative positional embedding and the attention mask for shifting operation are omitted in the figur e for simplicity. M: Window Size. n_w: Number of windows after partition. *opt.:* Optional Operation.

Setting of Temperature $\mathcal{T}$	mIoU(%)
Fixed value for all layers (Final Model)	57.57
Linear Decay	57.47
Learnable Parameters	57.58

Table A.1. Additional ablation study for different setting of $\mathcal{T}$ on the NYUv2 dataset.

Linear Decay. We explore the impact of linearly decreasing the temperature as the network goes deeper. This follows the intuition that the uncertain nodes are becoming more confident as they go through more layers and aggregate more information from other confident nodes. Specifically, we set the $\mathcal{T}_1 = 15$ and $\mathcal{T}_n = 5$, where n is the total number of UASA layers. The values of $\mathcal{T}$ in those layers in between are linearly interpolated. As we show in Table A.1, this yields slightly poorer performance than using the fixed value for all layers.

$\mathcal{T}$ as Learnable Parameters. Instead of manually finding $\mathcal{T}$ as a hyperparameter, we also explored to learn this value during end-to-end training. To keep the temperature within a reasonable range and avoid the difficulty of learning large values, we choose to learn an auxiliary parameter τ for each UASA layer, such that $\mathcal{T} = sigmoid(\tau) \cdot \mathcal{T}_{max}$. This formulation makes sure that $\mathcal{T}$ is within the range of $[0, \mathcal{T}_{max}]$, where we set $\mathcal{T}_{max} = 20$ and initialize $\tau = 0$. As shown in Table A.1, setting $\mathcal{T}$ to Learnable Parameters leads to slightly better per-

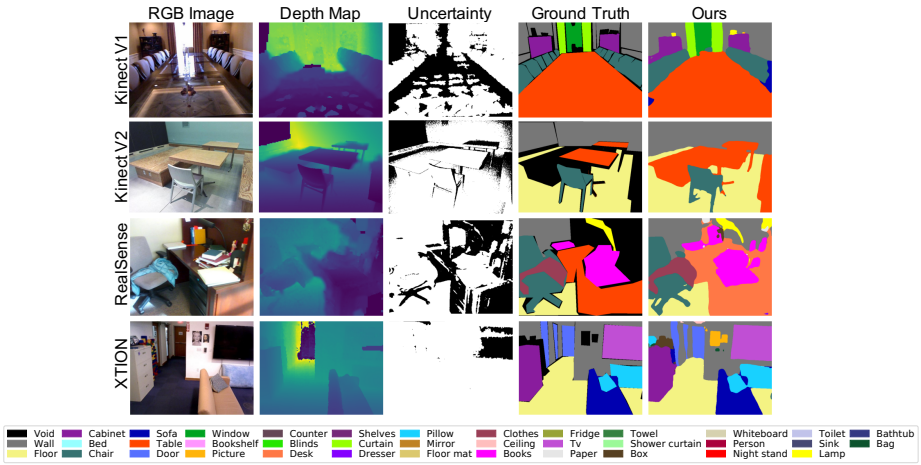

Fig. A.2. Additional Qualitative Results on SUN RGB-D benchmark. Samples in different rows are captured by different RGB-D sensors. Kinect: Microsoft Kinect. RealSense: Intel RealSense. XTION: ASUS XTION. Best view in color. Zoom in for more details.

formance but the difference is minor. Therefore, we still choose the fixed-value setting in our final model for simplicity. We believe that there might be a better way to learn this parameter which we plan to investigate in our future work.

A.3 Qualitative Results on SUN RGB-D

We provide qualitative results on SUN RGB-D dataset in Figure A.2. SUN RGB-D dataset consists of RGB-D images collected using four different types of sensors, and we visualize our results on samples collected by different sensors. We can see that our method produces stable and accurate results regardless of the sensor type, which demonstrates the generality of our approach.

References

1. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)